


Cross-language MeSH Indexing using Morpho-Semantic Normalization

Kornél Markó^a Philipp Daumke^a Stefan Schulz^a Udo Hahn^b

^aFreiburg University Hospital, Department of Medical Informatics (<http://www.imbi.uni-freiburg.de/medinf>)

^bFreiburg University,  Computational Linguistics Research Group (<http://www.coling.uni-freiburg.de>)

Abstract

We consider three alternative procedures for the automatic indexing of medical documents using MESH thesaurus identifiers as target units (document descriptors). Rather than considering complete words as the starting point of the indexing procedure, we here propose morphologically plausible subwords as basic units from which MESH terms are derived. We describe the morphological segmentation and normalization procedures, as well as the mappings from subwords to MESH terms, and discuss results from an evaluation carried out on a German-language corpus.

INTRODUCTION

Digital libraries, Internet portals and other digitally available collections of (medical) documents are not yet ready for the delivery of immediately useful health information. Often, relevant information has to be filtered out by browsing through directories or general purpose search engines must be used, which are not adapted to the special needs of medical information retrieval. The assignment of index terms has always been an important means to focus on relevant documents in large document collections. Adding this kind of meta information can either consist of identifying free-text keywords or assigning index terms out of a predefined and fixed set of descriptors. Using a thesaurus such as MESH [7, 3] or UMLS [8, 1], descriptors not only organize the document space in terms of semantically related groups, but also enhance search procedures by expanding queries by synonyms, more or less general terms, related terms, etc.

However, the manual assignment of index terms out of a (very large) set of descriptors is a highly cost-expensive expert task. In the biomedical domain, the most prominent example is the MEDLINE database maintained by the U.S. National Library of Medicine (NLM). Speeding up the indexing process by the use of automated document preprocessing in order to acquire candidate terms for the coders is a major desideratum of the NLM's indexing initiative (IND) [3]).

In this paper, we compare three automated indexing methods for medical documents, based upon morphological text analysis. The approach is language-independent and focuses on MESH terms as document descriptors.

MORPHOLOGICAL PROCESSING

Natural language is characterized by morphological processes, which tend to alter the literal appearance of text words but leave their core meaning by and large unchanged. Such morphological variants can generally be described as concatenations of basic lexical forms (stems) with additional substrings (affixes). We distinguish three kinds of morphological processes, viz. inflection (e.g., adding the plural *es* in “*leuk⊕o⊕cyt⊕es*”),¹ derivation (e.g., attaching the derivation suffix *ic* in “*leuk⊕o⊕cyt⊕ic*”), and composition (e.g., in “*leuk⊕em⊕ia*”).

Morphological analysis is concerned with the reverse processing, i.e., deflection (or *lemmatization*), derivation and decomposition. The goal is to map all occurring morphological variants to some canonical base form(s) — e.g., ‘*leuk*’ or ‘*cyt*’ in the examples from above. The diversity of morphological processes varies between languages, with English known as a morphologically “poor” language, while others, e.g. German, Italian or Russian are much more diverse.

In addition, medical terms are characterized by a typical mix of Latin and Greek roots with the corresponding host language, often referred to as *neo-classical compounding* [5], e.g., in words such as *neuroencephalomyelopathy*, *glucocorticoid*, *pseudohypoparathyroidism*. Morphologically rich languages (e.g., German) tend to conflate these terms, moreover, with host language terms, resulting in longer single-word compounds such as *Gastrointestinaltrakt*, *Kortikoidmedikation*, etc. Obviously, such phenomena have to be accounted for in a system that maps free-text expressions to an indexing vocabulary such as MeSH.

SUBWORD SEGMENTATION AND SEMANTIC NORMALIZATION

Our experimental setting combines heuristic and statistical matching procedures with the MORPHOSAURUS (an acronym for MORPHEME TheSAURUS) document pre-processing engine developed by the authors [9]. MORPHOSAURUS takes German and English ASCII and HTML texts as input and transforms them in three steps, viz. orthographic normalization, morphological segmentation and semantic normalization.

¹ ‘ \oplus ’ denotes the string concatenation operator.

Klinische Schwerpunkte stellen chronisch entzündliche Darmerkrankungen, die familiäre adenomatöse Polyposis, die akute Pankreatitis, die multimodale Therapie des Pankreaskarzinoms, sowie die Antibiotikatherapie sowohl prophylaktisch als auch bei Peritonitis dar.	klinische schwerpunkte stellen chronisch entzündliche darmerkrankungen die familiäre adenomatoese polyposis die akute pankreatitis die multimodale therapie des pankreaskarzinoms sowie die antibiotikatherapie sowohl prophylaktisch als auch bei peritonitis dar.	klin ische schwerpunkt e stell en chron isch entzuend liche darm erkrank ungen die famili aere adenom atoese polyp osis die akut e pankreat itis die multi modal e therap ie des pankreas karzinom s sowie die antibiotik a therap ie sowohl prophylakt isch als auch bei periton itis dar.	cliniijxqz focusiipwxk stelliipzip chronoiijrz itidesiixk splanchniirqp oticiiyii familiizxjr adeniiwqz oticiiyii polypipkw oticiiyii acutaiijz pancreatiqxir itidesiixk multiikrj modaliqxjr therapiipri pancreatiqxir oncoijwqj antibiosipyprw therapiipri prophylaktipkiw peritonikzqx itidesiixk.
--	---	---	---

Figure 1: Three steps to morphosemantic normalization of a document: The original document (box 1) is transformed orthographically (box 2) and then segmented according the subword thesaurus (box 3). Next, content bearing segments are mapped to MORPHOSAURUS equivalence classes, whose identifiers (MIDs) are automatically generated by the system (box 4).

Orthographic Normalization

A preprocessor unifies all characters from documents in lower-case 7-Bit ASCII and performs language-specific character substitutions (e.g., for German ‘ß’ → ‘ss’, ‘ä’ → ‘ae’, ‘ö’ → ‘oe’, ‘ü’ → ‘ue’). Additional translation rules are motivated by idiosyncrasies of the medical sublanguage: ‘ca’ → ‘ka’, ‘co’ → ‘ko’, ‘cu’ → ‘ku’, ‘ce’ → ‘ze’, ‘ci’ → ‘zi’, and others. This solves a notorious problem in German medical terminology where original Latin terms contain ‘c’ instead of ‘k’ and ‘z’, whereas German derivations of the same terms prohibit the use of ‘c’ — a rule frequently violated even by professional medical writers (e.g., the use of different surface forms such as “Karzinom”, “Karcinom”, “Carzinom”, “Carcinom”).

Morphological Segmentation

Based upon a German and English *subword lexicon*, the system segments each orthographically normalized input document into a sequence of semantically plausible sublexical units. Each document token t of length n defined as a sequence of characters c_1, c_2, \dots, c_n is processed in parallel by a forward and backward matching process. The forward matching process starts at position $k = n$ and decrements k iteratively by one unless the sequence c_1, c_2, \dots, c_k is found in the subword lexicon. Alternatively, the backward matching process starts at position $k = 1$ and increments k iteratively by one unless the sequence c_k, c_{k+1}, \dots, c_n is found in the lexicon. In each case, the substring found is entered into a chart. Now, unless the remaining sequences are not empty, $c_{k+1}, c_{k+2}, \dots, c_n$ and c_1, c_2, \dots, c_{k-1} , respectively, are tested recursively in the same manner, forward and backward. The segmentation results in the chart are checked for morphological plausibility using a finite-state automaton in order to reject invalid segmentations (e.g., segmentations without stems or beginnings with a suffix).

If there are ambiguous valid readings or incomplete segmentations (due to missing entries in the lexicon)² a series of heuristic rules are applied, preferring those segmentations with the longest match from the left, the lowest number of unspecified segments, etc.

²Currently, the MORPHOSAURUS subword lexicon contains approximately 20,000 entries for each German and English and approximately 12,000 entries for Portuguese.

Semantic Normalization

Using a *subword thesaurus* which basically defines intra- and interlingual equivalence classes, each semantically relevant sublexical unit produced by the morphological segmentation is replaced by its corresponding MORPHOSAURUS class identifier (MID, for details, cf. [9]).

Figure 1 illustrates the three procedures, *viz.* orthographic normalization, morphological segmentation and semantic normalization. The final result is a morphosemantic normalized document in a concept-like, language-independent target representation.

MAPPING PROCEDURES

In the following, we describe a heuristic, a statistical and a hybrid approach to automatically identify MESH [7] main headings as document descriptors. MESH, the NLM’s biomedical controlled vocabulary, consists of sets of terms denoting descriptors in a hierarchical structure. In the 2002 MESH, which we use, there are over 20,500 so-called main headings with over 120,000 synonyms (entries).³

Initially, for each of the methods, the texts to be indexed with MESH descriptors, as well as all English MESH main headings and (synonymous) entry terms undergo the morpho-semantic normalization procedure described in the previous section. The result is a *language independent* representation of both the (German) documents and the (English) indexing vocabulary in which words are substituted by their corresponding MIDs. This approach, in principle, allows processing documents in any language covered by MORPHOSAURUS.

Heuristic Approach

The first automated indexing method applies heuristic rules (some of them proposed by the indexing initiative (IND) of the NLM [3]) on a *normalized* text: First of all, every MESH descriptor whose normalized representation contains at least one of the MIDs in the document, is retrieved. Afterwards, each normalized MESH descriptor is evaluated against the normalized

³Publication Types, subheadings and the set of chemical supplementary terms are not considered in this work, whilst special descriptors such as *Age Groups*, *Check Tags* and *Geographics* are included.

$$w(MeSH_i | MID_1, \dots, MID_n) = \log \prod_{j=1}^{n-2} \begin{cases} \frac{P(MID_j, MID_{j+1}, MID_{j+2} | MeSH_i)}{P(MID_j, MID_{j+1}, MID_{j+2})} & , \text{if both nominator and denominator} > 0 \\ 1 & , \text{otherwise} \end{cases}$$

Figure 2: Formula for estimating the conditional weighting value w of a MESH descriptor i given n MORPHOSAURUS class identifiers (MIDs)

text by computing diverse factors. In this contribution, we confine ourselves to the most important metrics:

- **Longest Match Factor:** On the level of MIDs, individual MESH descriptors, which appear as single entries, can also appear together in additional MESH entries. For example, the German word “*Bauchschmerzen*” (“*abdominal pain*”) that appears in a text and is normalized to the MIDs “*abdomdiiiiiq*” and “*painiiijkj*” is, amongst others, associated to the MESH entries “Abdominal Pain” ([“*abdomdiiiiiq*”, “*painiiijkj*”]), “Abdomen” ([“*abdomdiiiiiq*”]) and “Pain” ([“*painiiijkj*”]). If two or more normalized MeSH descriptors can be merged to one longer MESH descriptor, it is preferred over the others.
- **Phrase Factor:** The number of different MIDs in a sentence participating to a normalized descriptor is called *MID number*. The *phrase interval* of a normalized descriptor, on the other hand, can be considered as the span between the first and the last MID associated with this descriptor within a sentence. The *phrase factor*, then, is defined as the ratio of *MID number* and *phrase interval*. As an example, the sentence “*die Leber des Patienten wurde transplantiert*” (“*the patient’s liver was transplanted*”) transforms to [“*hepatiiipji*”, “*patientiiikzix*”, “*transplantiiiqjxw*”]: *MID number*(“*hepatiiipji*”, “*transplantiiiqjxw*”) = 2; *phrase interval* = 3 ; *phrase factor* = 2/3.
- **Entry Factor:** The *entry factor* is the *MID number* divided by the number of morphemes of the associated descriptor. For example, the German noun phrase “*nodulaere Hyperplasie*” (“*nodular hyperplasia*”) is normalized to [“*noduliikwrk*”, “*abovei-iiijy*”, “*plastiipixi*”] and the MESH descriptor “Focal Nodular Hyperplasia” to [“*focaliizy*”, “*noduliikwrk*”, “*abovei-iiijy*”, “*plastiipixi*”]: *entry factor* = 3/4.
- **Title Factor:** A descriptor found in the title will be ranked higher than others.

Finally, all possible descriptors are ordered according to a weighted average of the above (and other) metrics.

Statistical Approach

We start here from a larger collection of MEDLINE abstracts that have MESH main headings already assigned. Based on this data, we pursue a modified Bayesian approach that computes statistical evidence for MORPHOSAURUS class identifier (MID) trigrams by counting their frequency of appearance in the training corpus, subject to the actually occurring annotated MESH entries.

In the test phase, when we aim at extracting MESH terms as valid descriptors for a document, we rank these terms by their *weighting* (w) values as defined in Figure 2. Given a document which contains n class identifiers, the conditional weighting value for a MESH main heading $MeSH_i$ is computed by the product of the conditional probabilities P of the MID trigrams in the text that co-occur with the descriptor $MeSH_i$ in the training set, divided by the plain probability of the corresponding text trigrams in the training collection – if both probabilities are observable, at all. Here, the denominator takes into account the fact that infrequent terms have a greater explanatory power for a given entity when faced with large quantities of data and, hence, increase the weighting value for that entity. If no trigram that is currently being processed appears in the training data, or if it is not associated with the current descriptor $MeSH_i$, that subresult is simply multiplied with 1. This expresses the fact that there is no evidence for a further refinement – simply because the observation that the combination of a trigram and a MESH descriptor does not appear in the training set does not mean that it may never occur, at all.

In our approach MID trigrams are treated in an unordered way. They are defined as a set of MIDs that co-occur within a document window of three items, regardless of the original sequence of words that produced the set of MIDs. The reason for this is that in German, as well as in English, the MID order changes when genitives or prepositions come into play, as with “*femoral neck fracture*” vs. “*fractured neck of femur*” corresponding to [“*femuriiizir*”, “*nuchaliijkwq*”, “*fracturiiizzx*”] vs. [“*fracturiiizzx*”, “*nuchaliijkwq*”, “*femuriiizir*”].

The method described here substantially contrasts with the statistical approach that was chosen by the indexing initiative from NLM [3], since they considered plain *character* trigrams in their studies.

Hybrid Approach

A combination of the heuristic and the statistical approach is achieved in two steps. First, all descriptors that are ranked in the top 30 by both of the methods are set to the top of the resulting list. In a second step, two entries on the top of the output of the statistical approach are alternately incorporated into the final result, followed by one entry of the heuristic approach. Previous experiments have shown that this empirically motivated procedure leads to better results than a more formal one, e.g., by simply multiplying the outcome values of the different weighting functions.

EVALUATION

For our evaluation we decided to use abstracts of German medical journal publications, available by an online library for medicine, the “*SpringerLink*”,⁴ which contains, among others, dozens of medical journals. We chose those journals covering clinical disciplines and which were indexed by MEDLINE. The relevant MESH identifiers were extracted from PUBMED⁵.

Then we randomly defined distinct text collections for the training phase (914 abstracts - 161,158 words for the heuristic, and 4,048 abstracts - 699,144 words for the statistical approach) and the test phase (309 abstracts - 50,161 words). The abstracts were processed according to the methods described in the previous section and the classification results are evaluated against the supplied MESH main headings, which - equivalent to the study of the indexing initiative of the NLM [2] - serve as the *de facto* gold standard for our experiments.⁶

For the mapping experiments we distinguished the three different conditions, viz. *heuristic* mapping, *statistical* mapping, and a *combination* of both. As common in information retrieval experiments, we focus on resulting precision/recall values.

Evaluation Results

Table 1 depicts the values for precision and recall for the chosen test scenarios. For each of the three methods we considered the top 5, 10 and 40 ranked descriptors. The measurements we use here were introduced

⁴<http://link.springer-ny.com/>

⁵<http://www.ncbi.nlm.nih.gov/PubMed/>
The *Check Tags* “*English Abstract*” and “*Human*” are excluded in this study, since they appear in almost every document.

⁶Unfortunately, since these encodings are done by hand by the editors of NLM, the types of information that were added with the descriptors varied from one document to another. The MESH term “*Germany*”, for example, can serve as a document descriptor in almost every document that refers to a German hospital, a German clinical study, etc. In some cases, this entry was assigned to a document - in others it was not. Such inconsistencies in the test collection will affect the quality of evaluation results when we take this data as gold standard.

Cut-Off	5		10		40	
P/R	P	R	P	R	P	R
Heur.	38.9	18.1	28.9	27.0	9.6	35.5
Stat.	48.6	23.3	36.8	34.7	16.7	61.4
Comb.	54.5	25.9	41.5	39.1	17.6	65.0

Table 1: Evaluation Results — Precision/Recall (P/R) Table for the different approaches at different cut-off points.

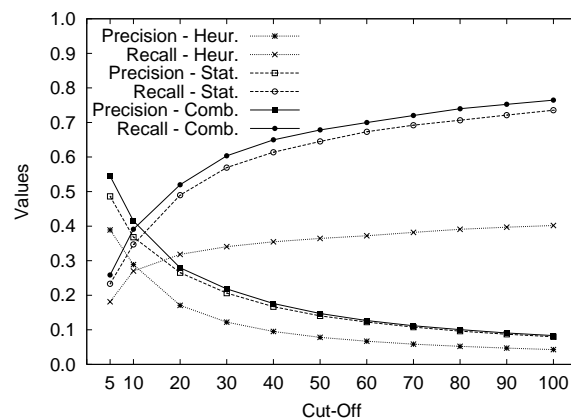


Figure 3: Evaluation Results — Precision/Recall graphs for the heuristic, statistical and combined approaches at different cut-off points.

by [2] and we include them in order to compare the results more accurately.

Considering the set of five chosen descriptors, the heuristic method retrieves 18% of all relevant terms at a precision rate of 39%. The statistical approach performs significantly better (49% precision, 23% recall) whilst the combined method reaches 55% precision at a recall of 26%. The differences between the described methods become more significant when we regard the top 40 of the system’s proposed descriptors. Whereas precision values accumulate at 10%, 17% and 18% for each of the methods, recall increases from 36% for the heuristic approach up to 61% for the statistical method and reaches at 65% for the combined algorithm. Ignoring MESH’s *Check Tags* and *Age Groups*, which tend to be easier to identify, our combined mapping procedure still reaches 43% precision at a recall rate of 28% (top 5) and 12% precision at 61% recall (top 40), respectively.

Summarizing, Figure 3 shows the resulting precision/recall value pairs for the different methods for the top 5, 10, 20, etc. up to the top 100 proposed descriptors (including *Check Tags*). The crossings of the lines in the figure indicate that the abstracts of the test collection are predominantly assigned to more than ten descriptors.

RELATED WORK

The work of Lovis *et al.* [4] and Zweigenbaum *et al.* [10] reveal the usefulness of morphological knowledge for automatic indexing, at least for French as a morphologically rich language. For German, however, the proposed method, *viz.* the enumeration of morphological variants in a semi-automatically generated lexicon (also cf.[1]), turns out to be infeasible, since the German language is morphologically extremely productive.

In direct comparison to the system that is proposed by the indexing initiative (IND) from the NLM [2, 3], which reaches 60% precision at a recall level of 29% (top 5) and 20% precision at 61% recall (top 40) using their most favored combined method "MetaMap Indexing" with "PubMed Related Citations", our combined method shows lower performance regarding the precision values. Nevertheless, considering the top 40, our approach retrieves slightly more relevant descriptors (65% vs. 61%).

The loss of performance in this comparison (i.e., in fact not only a comparison of the plain methods, but also a cross-language comparison) can be interpreted as a direct consequence from the language-specific morphological complexity inherent to German.

The indexing system presented in [6] dealing with documents on high energy physics reaches 60% both for precision and recall. This superior performance can certainly be ascribed to the use of entries of the limited DESY thesaurus⁷ (approx. 2,700 entries - compared to over 20,500 MESH terms). In contrast, medical language – in comparison to the narrowed and more precise domain terminology of physicists – certainly produces more lexical variants with reference to the morphological processes taken into account in this contribution.

CONCLUSION

We presented promising approaches to machine supported indexing of biomedical texts using entries from the MESH thesaurus. Our system is based upon sub-lexical units and a language-independent thesaurus to represent content bearing elements of a document. This normalized representation, together with heuristic and statistical procedures, lead to quite accurate document indexes. A possible application scenario could be the support of indexers in case of facing non-English documents that lack translated abstracts.

Future investigations will emphasize the multilingual aspect of our approach. In this regard, evaluations for the currently supported languages, German, English and Portuguese, are planned while the methods for automatized indexing have to be further refined.

Acknowledgments. This work has been supported by grant Klar 640/5-1 from the *Deutsche Forschungsgemeinschaft (DFG)*.

References

- [1] A. R. Aronson. Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. In *AMIA 2001 – Proc. of the Symposium of the American Medical Informatics Association*, 2001.
- [2] A. R. Aronson, O. Bodenreider, and H. F. Chang et al. The indexing initiative. In *A Report to the Board of Scientific Counselors of the Lister Hill National Center for Biomedical Communications*, 1999.
- [3] A. R. Aronson, O. Bodenreider, and H. F. Chang et al. The NLM indexing initiative. In *AMIA 2000 – Proc. of the Annual Fall Symposium of the American Medical Informatics Association*, pages 17–21, 2000.
- [4] Christian Lovis, Pierre-André Michel, Robert H. Baud, and Jean-Raoul Scherrer. Word segmentation processing: A way to exponentially extend medical dictionaries. In R. A. Greenes, H. E. Peterson, and D. J. Protti, editors, *MEDINFO'95 – Proceedings of the 8th Conference on Medical Informatics*, number 8 in IFIP World Conference Series on Medical Informatics, pages 28–32. Vancouver, Canada, 1995. Amsterdam: North-Holland, 1995.
- [5] Alexa T. McCray, A. C. Browne, and D. L. Moore. The semantic structure of neo-classical compounds. In R. A. Greenes, editor, *SCAMC'88 – Proceedings of the 12th Annual Symposium on Computer Applications in Medical Care*, pages 165–168. Washington, D.C.: IEEE Computer Society Press, 1988.
- [6] A. Montejó Ráez. Towards conceptual indexing using automatic assignment of descriptors. In *Workshop in Personalization Techniques in Electronic Publishing on the Web: Trends and Perspectives. Málaga, Spain*, 2002.
- [7] NLM. *Medical Subject Headings*. Bethesda, MD: National Library of Medicine, 2001.
- [8] NLM. *Unified Medical Language System*. Bethesda, MD: National Library of Medicine, 2002.
- [9] S. Schulz and U. Hahn. Morpheme-based, cross-lingual indexing for medical document retrieval. In *International Journal of Medical Informatics*, 59(3), pages 87–99, 2000.
- [10] P. Zweigenbaum, S. Darmoni, and N. Grabar. The contribution of morphological knowledge to French MESH mapping for information retrieval. In *AMIA 2001 – Proc. of the Symposium of the American Medical Informatics Association*, pages 796–800, 2001.

⁷DESY. The high energy physics index keywords, 1996 (<http://www-library.desy.de/schlagw2.html>)